

ИНТЕЛЛЕКТУАЛЬНАЯ СИСТЕМА КЛАССИФИКАЦИОННОГО ПРОГНОЗИРОВАНИЯ УСПЕШНОСТИ СОЦИАЛЬНЫХ ПРОЕКТОВ

О. В. Федоришин, В. Д. Карпуша, А.М. Телиженко
Сумский государственный университет, г. Сумы

В статье предложен научно-методический подход к оценке эффективности реализации, успешности внедрения, целесообразности реорганизации, актуальности слияния социально-ориентированных кампаний, стартапов, проектов, программ. Рассмотрены основные аспекты формирования входного математического описания системы, особенности ее функционирования в режиме обучения и экзамена, а также основные критерии оценки информационной эффективности интеллектуальной технологии. Выполнено классификационное прогнозирование успешности страховых стартапов на основе применения парадигмы искусственных нейронных сетей по алгоритму обратного распространения ошибки. Для повышения достоверности классификации выполнена оптимизация структуры нейронной сети.

Ключевые слова: классификационное прогнозирование, успешность стартапа, страховая компания, входное математическое описание, искусственные нейронные сети, алгоритм обратного распространения ошибки, многослойный перцептрон.

ВВЕДЕНИЕ

Современные информационные системы анализа данных страховых компаний содержат в своем составе интеллектуальные элементы, позволяющие оценивать потенциальные возможности клиентов принимать участие в различных программах страхования и предоставления новых страховых продуктов. Использование интеллектуальных систем позволяет классифицировать клиентов по определенным характеристикам, например, степени их доверия к страховой компании, определять их заинтересованность в увеличении объема страховых услуг. Интеллектуальный анализ данных (Data Mining) и классификационное прогнозирование являются распространенными инструментами интеллектуальных информационных технологий и содержит как классические технологии экспертных знаний, так и современные разработки в области нечетких множеств и искусственных нейронных сетей. Формирование баз знаний и использование классифицированных данных, позволяет еще на этапе планирования оценить степень успешности проекта и своевременно внести изменения в его структуру.

В статье предложен подход к созданию интеллектуальной системы прогнозирования успешных стартапов на примере интеллектуального анализа данных страховой компании с элементами искусственного интеллекта для оценки активности клиентов компании в специальных программах страхования «Mobile Home Policy».

ТЕХНОЛОГИЯ ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА ИНФОРМАЦИИ. ПОСТАНОВКА ЗАДАЧИ.

Использование методов машинного обучения и теории распознавания образов [1] позволяет применить интеллектуальные системы [2], способные анализировать данные подобно человеку [3]. Современное состояние уровня автоматизации украинских предприятий находится на переходном этапе [4]. При этом, применение Data Mining [5], или knowledge discovery является одним из подходов, который используется для первоначального анализа данных. Использование данного подхода

позволяет выделять информацию от данных, выявлять закономерности, сходство в распределении и т.д.. Data Mining является одним из разновидностей низкоуровневого анализа данных универсальной природы, что делает его незаменимым для различных прикладных задач [6].

Целью статьи является обоснование информационного и программного обеспечения интеллектуальной системы прогностической классификации активности клиентов страховой компании в специальной программе страхования «Mobile Home Policy» на основе анализа их социодемографических характеристик и текущих страховых обязательств. В статье рассматриваются вопросы:

- формирования входного математического описания интеллектуальной системы;
- выбора типа и структуры искусственной нейронной сети, обучаемой «с учителем»;
- разработки и реализации алгоритмов оптимизации функциональных параметров нейронной сети;
- оценки воздействия на достоверность классификации такого параметра обучения, как передаточная функция персептронов искусственных нейронных сетей (ИНС);
- проверки работоспособности блока интеллектуальной системы на примере задачи определения потенциальных участников специальной программы страхования «Mobile Home Policy».

АЛГОРИТМ ОБУЧЕНИЯ ИНТЕЛЛЕКТУАЛЬНОЙ СИСТЕМЫ

Для реализации интеллектуальной системы прогнозирования избрана нейросетевая парадигма, а именно многослойный персептрон Розенблатта [7]. Нейросеть с данной топологией используется для решения ряда прикладных задач типа Data Mining [8]. Кроме того, возможно использование оптимально сконфигурированной нейросети для оценки состояний исследуемого объекта. Для повышения достоверности распознавания целесообразно использовать алгоритм обратного распространения ошибки [9]. Новизной применения ИНС такой топологии является оптимизация передаточной функции [10], т.е. функционала, который на основе значений входных синапсов (весов) нейрона, формирует его выходной сигнал. Для многослойного персептрона выходной сигнал внутреннего слоя нейронов представляет собой входные данные для нейронов внешнего уровня.

Синтез ИНС, кроме формирования ее структуры, также заключается в обучении. Обучение представляет собой конфигурирование нейронов, путем поочередной модификации веса синапсов. Период корректировки всех нейронов нейросети называют эпохой или циклом обучения. Перед началом обучения необходимо, прежде всего, вес синапсов инициализировать начальными значениями весов. При введении запоминающего стимула (результат классификации) происходит реакция синапсов внешнего слоя, т.е. формируется выходной сигнал, значение которого следует трактовать как результат классификации. Когда образованная волна достигает внешних нейронов, можно определить величину ошибки как разницу между полученным и желаемым значением реакции нейросети [11]. Ошибка классификации является основным аргументом оценки достоверности классификации данных с помощью нейросети.

Предложенный алгоритм использует определенное внешнее звено, – целевые выходные образы, которые создаются экспертамина предыдущем этапе для каждого входного образа. Такие алгоритмы называются «алгоритмами обучения с учителем» [12].

Оценка реакции скрытых нейронов осуществляется вычислением взвешенного значения ошибки классификации для слоя нейронов. В качестве весовой функции используют текущие значения синапсов нейронов наружного слоя, которые идут от

скрытых нейронов к эффекторам выходного слоя. Ошибка распространяется в обратном направлении, корректируя синапсы. Если скрытых слоев несколько, оценку ошибок приводят для каждого, начиная с слоя эффекторов.

Целевая функция ошибки нейросети, которая должна быть минимизирована, представляет собой:

$$E(w) = \frac{1}{2} \sum_{j,p} (y_j^{(N)} - d_{j,p})^2, \quad (1)$$

где $y_{j,p}^{(N)}$ – реальное исходное состояние нейрона j выходного слоя N ИНС при подаче на ее входы p -ого образа; $d_{j,p}$ – идеальное (целевое) состояние этого нейрона.

Итог ведется по нейронам выходного слоя и по всем обрабатываемым ИНС образам. Минимизация осуществляется по методу градиентного спуска, то есть веса модифицируются следующим образом:

$$\Delta w_{ij}^{(n)} = -\eta \cdot \frac{\partial E}{\partial w_{ij}}, \quad (2)$$

где w_{ij} – весовой коэффициент синапса, соединяющего i -тый нейрон слоя $n-1$ с j -тым нейроном слоя p , η – коэффициент скорости обучения ($0 < \eta < 1$).

Для выходного слоя:

$$\delta_j^{(N)} = (y_j^{(N)} - d_j) \cdot \frac{dy_j}{ds_j} \quad (3)$$

Для всех остальных слоев:

$$\delta_j^{(n)} = \left[\sum_k \delta_k^{(n+1)} \cdot w_{jk}^{(n+1)} \right] \cdot \frac{dy_j}{ds_j} \quad (4)$$

РЕЗУЛЬТАТЫ МОДЕЛИРОВАНИЯ

В качестве примера применения интеллектуальной системы рассмотрено прогнозирование активности клиентов компании по участию в специальной программе страхования «Mobile Home Policy». Формирование входного математического описания для интеллектуальной системы выполнялось по данным размещенным на репозитории данных для машинного обучения (Machine Learning Repository) Центра машинного обучения и интеллектуальных систем университета Калифорнии.

В работе применяются данные раздела Insurance Company Benchmark (COIL 2000) Data Set. Обучающая выборка в виде матриц типа «объект – свойство» формируется для двух классов клиентов компании: принимающих участие в специальных программах страхования, и тех, которые отказываются от участия. Словарь признаков состоит из 85 характеристик клиентов, например, возраст, должность,

семейное положение, наличие других страховых полисов и т.п.. Обучающая выборка состояла из 4000 реализаций, из которых 238 отвечали положительно на предложение об участии в специальной программе.

Программная реализация системы выполнялась в среде MATLAB с использованием специализированного пакета для формирования, обучения и тестирования искусственных нейронных сетей Neuro Net Toolbox.

На рис. 1 приведен график динамики изменения значения среднеквадратической ошибки в процессе обучения нейросети с передаточной функцией нейронов скрытого и выходного слоя в виде логарифмической функции:

$$f(s) = \log \text{sig}(s) = \frac{1}{1 + e^{-s}} \quad (5)$$

где s – результат работы сумматора искусственного нейрона.

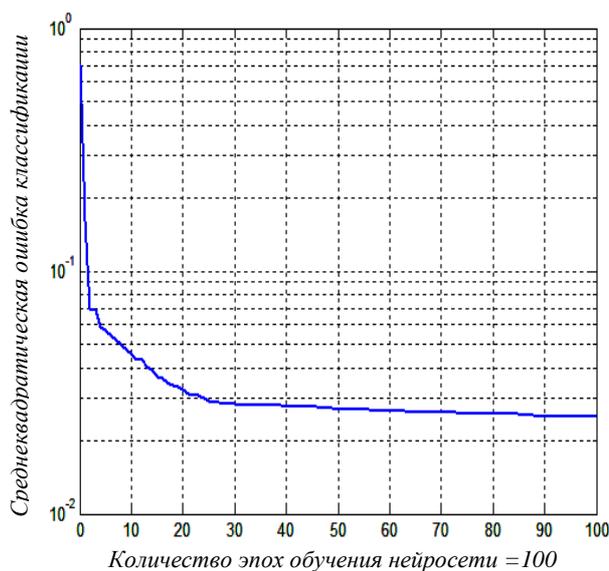


Рис . 1 - Динамика изменения значения среднеквадратической ошибки при обучении ИНС с логарифмической передаточной функцией.

Анализ рис.1 показывает, что нейронной сети не удалось построить безошибочный классификатор, поскольку значение ошибки больше нуля ($E=0.0249859$).

Вероятность правильного определения клиентов, реакция которых на рекламную кампанию будет положительной составляет $D1=99.73\%$, а вероятность правильной классификации тех, кто не примет участие – $D2=59.66\%$. С целью повышения этих значений был предложен m - сценарий, в котором изменялся такой важный параметр как тип передаточной функции. При этом изменялась передаточная функция нейронов скрытого и выходного слоя на линейную:

$$f(s) = \text{purelin}(s) = s \quad (6)$$

где s – результат работы сумматора искусственного нейрона.

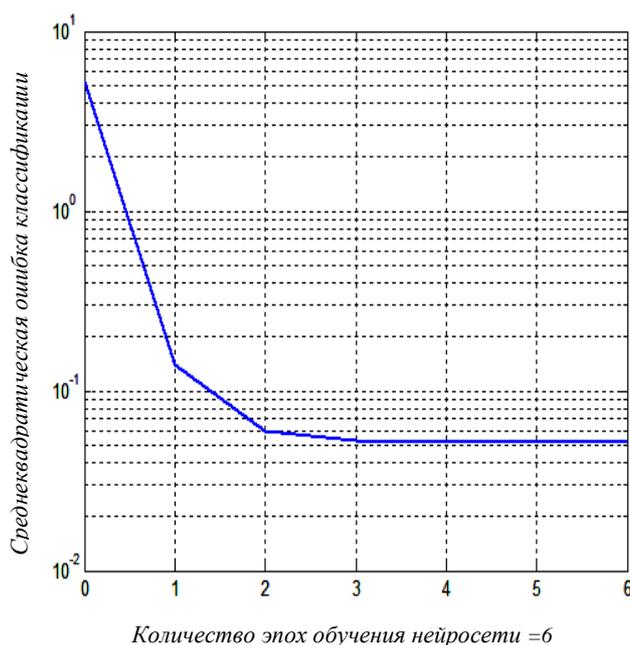


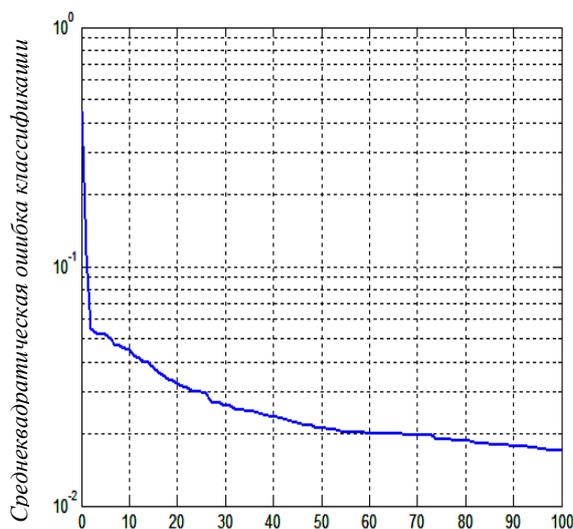
Рис. 2 - Динамика изменения значения среднеквадратической ошибки при обучении ИНС с линейной передаточной функцией.

Анализ рис. 2 показывает, что значение среднеквадратичной ошибки составляет $E=0.052218$. При этом точность распознавания реализаций первого класса $D1=99.21\%$, а для второго – $D2 = 0.97\%$. Это доказывает то, что классы клиентов являются линейно-неразделимыми. Следовательно, такой вид передаточной функции для данной системы применять нельзя.

Наиболее эффективным было применение передаточной функции в виде гиперболического тангенса:

$$f(s) = \text{tansig}(s) = \frac{2}{1 + e^{-2s}} - 1 \quad (7)$$

где s – результат работы сумматора искусственного нейрона.



Количество эпох обучения нейросети = 100

Рис. 3 - Динамика изменения значения среднеквадратической ошибки при обучении ИНС с передаточной функцией в виде гиперболического тангенса.

В данном случае, достоверность решений относительно первого класса клиентов составляет $D_1=98.37\%$, а второго класса – $D_2=75.14\%$. Значение среднеквадратической ошибки составляет $E = 0.0169204$.

Аналогичным образом выполнялось переобучение искусственной нейросети для различных вариантов передаточных функций скрытого и выходного слоя. Результаты переобучения приведены в табл. 1.

Таблица 1

Выбор передаточных функций нейросети

		Передаточные функции выходного слоя					
		$\tan sig(s)$		$purelin(s)$		$\log sig(s)$	
Передаточные функции скрытого слоя	$\tan sig(s)$	$E=0.0169$		$E=0.0464$		$E=0.0465$	
		$D_1=98.37\%$	$D_2=75.14\%$	$D_1=99.10\%$	$D_2=20.17\%$	$D_1=99.95\%$	$D_2=21.01\%$
	$purelin(s)$	$E=0.0522$		$E=0.0522$		$E=0.0462$	
		$D_1=99.21\%$	$D_2=0.97\%$	$D_1=99.21\%$	$D_2=0.97\%$	$D_1=99.84\%$	$D_2=23.95\%$
	$\log sig(s)$	$E=0.0179$		$E=0.0306$		$E=0.0250$	
		$D_1=99.89\%$	$D_2=73.11\%$	$D_1=99.87\%$	$D_2=42.02\%$	$D_1=99.73\%$	$D_2=59.66\%$

Анализ табл. 1 показывает, что оптимальной, в информационном смысле, является нейросеть, использующая передаточные функции в виде гиперболического тангенса.

ВЫВОДЫ

Практическую значимость полученных результатов свидетельствует о том, что предложенный алгоритм может быть внедрен в существующие комплексы для классификации и анализа данных. Его целесообразно применять для оценки социальных, экономических, политических программ и стартапов, оценивать выход на рынок новых продуктов и качество ребрендинга существующих и т.д.. Поскольку достоверность классификации не является безошибочной, для повышения качества классификации необходимо оптимизировать структуру нейросети путем увеличения количества входящих нейронов (признаков распознавания) и их оптимизации, увеличения количества скрытых слоев нейронов, оптимизировать передаточную функцию отдельно для каждого нейрона, а не для целого слоя, разработать гибридный алгоритм обучения искусственных нейронных сетей с использованием подходов и принципов других технологий машинного обучения и распознавания образов.

SUMMARY

INTELLIGENCE SYSTEM OF CLASSIFICATION FORECASTING OF SOCIAL PROJECTS SUCCESSFULNESS

O. Fedoryshyn, V. Karpusha, O. Telizhenko

Sumy State University, Sumy

The article presents the scientific and methodical approach of performance evaluation of implementation, application success, feasibility of reorganization, merging relevance of social oriented companies, startups, projects, programs etc. The intelligence system of startups classification according to the level of success is implemented. Herewith the basic aspects of formation of the input mathematical representation of a system, specifics of its functioning in mode of study and examination and basic criteria of performance evaluation of intelligent technologies in information conception are considered. The classification forecasting of the successfulness of insurance startups was carried out on the basis of application of the paradigm of artificial neural networks under the back propagation of error algorithm. To improve the reliability of the classification the structure of neural network is optimized.

Keywords: *classification forecasting, startup successfulness, the insurance company, input mathematical representation, artificial neural network, back propagation of error algorithm, multilayer perceptron.*

СПИСОК ЛІТЕРАТУРИ

1. Bishop С.М. «Neural Networks and Pattern Recognition». Oxford Press. 1995.
2. Башмаков А. И., Башмаков И. А. Интеллектуальные информационные технологии: Учеб. пособие. — М.: Изд-во МГТУ им. Н. Э. Баумана, 2005. — 304 с
3. Калініна І.О. Дослідження нейромережових методів у задачах прогнозування // Наукові праці. Вип.93. Т.106. — К., 2009
4. Береза А. М. «Інформаційні системи і технології в економіці» // Навчально-методичний посібник для самостійного вивчення // 2002 рік, 278 с.

5. Дюк В., Самойленко А., «Data Mining. Учебный курс» Спб: Питер, 2001, 386 с
6. Ian H. Witten, Eibe Frank and Mark A. Hall Data Mining: Practical Machine Learning Tools and Techniques. — 3rd Edition. — Morgan Kaufmann, 2011. — P. 664.
7. Соколов Е.Н., Вайтнявичус Г.Г. Нейроинтеллект: от нейрона к нейрокомпьютеру. – М.: Наука, 1989. – С. 283.
8. Уоссермен Ф. Нейрокомпьютерная техника – М.: Мир, 1992.
9. Kohonen T. «Self-organization and Associative Memory», Berlin: Springer-Verlag, 1989.
10. Minsky M. L, Papert S. 1969. Perceptrons. Cambridge, MA: MIT Press. (Русский перевод: Минский М. Л., Пейперт С. Перцептроны. – М: Мир. – 1971.)
11. Fausett L.V. «Fundamentals of Neural Networks: Architectures, Algorithms and Applications», Prentice Hall, 1994.
12. Hecht-Nielsen, Robert, «Counter-Propagation Networks», IEEE First International Conference on Neural Networks, Volume II, 1987.